VISIT…

LANZAROTE Caliente.COM

Материалы XIII Всеросс. семинара «Нейроинформатика и её приложения», Красноярск, 2004. 196с. – С.152-157.

ОПТИМИЗАЦИЯ ЭКСПЕРТОВ BOOSTING-КОЛЛЕКТИВА ПО ИХ КРИВЫМ ОБУЧЕНИЯ

Царегородцев В.Г.
www.NeuroPro.ru tsar@neuropro.ru

Практически исследована возможность оптимизации каждого эксперта boosting-коллектива на основе построения и анализа его “кривых” обучения – зависимостей ошибок обучения и обобщения от размеров обучающей выборки и свойств модели. Экстремальные свойства кривых обучения проявляются четко, различия кривых у экспертов подтверждают различия свойств специфически формируемых для каждого эксперта обучающих выборок, что позволяет оптимально настроить каждого эксперта на его выборку и в итоге максимизировать прогнозные способности всего коллектива.

1. Введение

Если одна модель недостаточно точно решает задачу прогнозирования или классификации с учителем, то несколько подобных моделей можно объединить в коллектив путем усреднения их прогнозов или голосования – при этом точность коллективного решения превзойдет точность отдельных моделей-экспертов. Эта идея нашла свое воплощение в нескольких методах – один из них, а именно boosting-алгоритм [1], здесь и рассматривается. Он предполагает последовательное построение моделей и обучение каждой последующей модели на выборке, в которой повышается доля примеров, неправильно решенных сформированным из ранее построенных моделей коллективом. Варианты алгоритма, например, с динамической адаптацией способа голосования [2], здесь рассматривать не будем.

В процессе построения boosting-коллектива происходит фокусировка обучения на все более и более трудных примерах или примерах, касательно которых ранее построенные эксперты дают противоречивое мнение, поэтому обучающая выборка для построения каждой очередной модели-эксперта все сильнее перестает отражать генеральную совокупность, описываемую исходной выборкой данных. Именно эта идея позволяет добиться малой скоррелированности ошибок отдельных моделей и повысить точность коллектива по сравнению с точностью отдельной модели.

Однако, кажется перспективной специальная оптимизация каждой модели: при смене законов распределения признаков (в искусственно формируемых выборках по сравнению с исходной) должна переоптимизироваться предобработка данных [3], подтверждаться оптимальность размера каждой обучающей выборки для адекватного оценивания параметров модели фиксированной структуры (сложности) [4], либо дополнительно должна оптимизироваться еще и сложность модели для максимизации обобщающих способностей последней [5]. Идеи К.Кортес с соавторами [4,5] о "кривых обучения", т.е. зависимостях ошибок обучения и обобщения от размера обучающей выборки или информационной емкости модели, и применяются в работе для выбора оптимальных настроек. В качестве класса моделей взяты обучаемые с учителем искусственные нейронные сети обратного распространения ошибки [6].

Эксперименты проведены на основе базы реальных данных CoverType из UCI KDD Database Repository (http://kdd.ics.uci.edu/) и показывают четкое проявление экстремальных свойств кривых обучения, что позволяет целенаправленно оптимизировать каждую очередную модель коллектива. Для баз данных, объема которых не хватает для построения нужного числа моделей, возможно зашумление данных или применение деформирующих исходные данные преобразований (см.

пример [7] для задачи распознавания рукописных символов) для получения новых примеров, добавляемых к исходной выборке для увеличения её объема.

2. Описание идеи boosting-алгоритма формирования коллектива экспертов

Классический boosting-алгоритм [1] выглядит следующим образом. Начальный "эксперт" обучается на некотором фрагменте большой исходной выборки данных. Далее из оставшейся части исходной выборки формируется выборка для обучения второго эксперта (равная по размерам выборке, использованной при обучении первого), половина которой состоит из примеров, неправильно решенных первым экспертом, а половина – из правильно решенных первым экспертом примеров. Т.о., для обучения второго эксперта повышается относительный вес примеров, трудных для решения первому эксперту, по сравнению с их распределением в генеральной совокупности, при требовании, что первый эксперт правильно распознает более 50% примеров выборки. Выборка для третьего эксперта формируется из тех не использованных при обучении первых двух экспертов примеров, касательно которых у этих двух экспертов существует противоречивое мнение.

При реальной же работе такого коллектива из трех экспертов новый вектор значений независимых признаков предъявляется первым двум экспертам, и если прогноз обоих одинаков, то этот прогноз и принимается. В случае расхождения прогнозов принимается прогноз третьего эксперта.

Из описания способа построения коллектива видно, что каждый из экспертов обучается на выборке одинакового и заранее заданного размера, что представляется неоптимальным. Именно оптимизацию размера выборок для каждого шага и рассмотрим здесь, запомнив вдобавок и возможность оптимизации сложности (информационной ёмкости) каждой модели по аналогичным принципам.

3. "Кривые" обучения (learning curves)

Ошибкой обучения назовём достигнутую после обучения точность решения примеров обучащей выборки, ошибкой обобщения – точность решения примеров независимой тестовой выборки. Идеализированные зависимости ошибок обучения и обобщения от размера обучающей выборки и от сложности модели, представленные на Рис.1-2, известны в теории машинного обучения под названием "learning curves" давно и считаются очевидными и общепринятыми – в учебниках (например, в [8]), воспроизводящих идеализированных "кривые" обучения, ссылок на какой-либо из источников не приводится. Реальные "кривые" обучения представлены в большом числе публикаций – в нейроинформатике в дополнение к [4,5] можно назвать более раннюю работу [9]. Ценность же [4,5] заключается в выдвижении и подтверждении общей гипотезы о том, что, получив несколько экспериментальных точек на такой кривой, последнюю можно аппроксимировать достаточно простым законом и далее интерполировать и экстраполировать зависимость для предсказания эффекта от того или иного размера выборки или сложности модели. На основе идей [4,5] можно получить и оценку уровня шума или уровня противоречивости выборки, ниже которого среднюю ошибку решения новых примеров снизить не удастся. В [10] показано, что из нескольких интер- и экстраполирующих зависимостей одна стабильно является наиболее точной на наборе рассмотренных в [10] баз данных, что подтверждает утверждения [4,5] об универсальности подхода и возможностях прогноза.

Рис.1 показывает выход на асимптоту, соответствующую внутреннему уровню шума или противоречивости задачи, с ростом обучающей выборки до размера, с которого запоминание обучающей выборки за счет снижения качества обобщения становится невозможной. Рис.2 показывает проявление меморизации обучающей выборки и ухудшения обобщения начиная с превышения оптимальной сложности

модели, причем модель с меньшей информационной ёмкостью тоже неоптимальна, поскольку не может аппроксимировать все существенные зависимости в данных.

Наиболее известная “кривая обучения” получается при замене на Рис.2 оси сложности модели на ось числа эпох обучения нейросети, но является вторичной по отношению к кривым Рис.1-2, поскольку расхождение ошибок обучения и обобщения при длительном обучении нейросети возможно только при недостаточном объеме обучающей выборки или при избыточной информационной ёмкости модели.

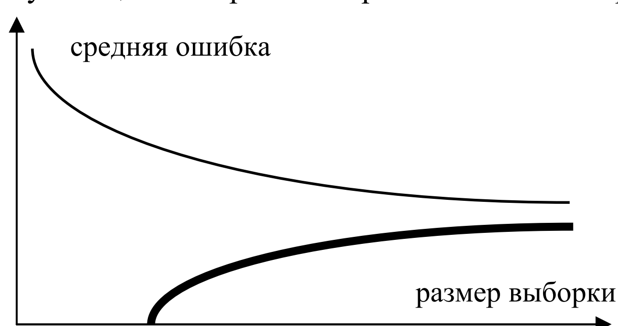


Рис.1. Теоретическая зависимость ошибок обучения (толстая линия) и обобщения (тонкая линия) от размера обучающей выборки.



Рис.2. Теоретическая зависимость ошибок обучения (толстая линия) и обобщения (тонкая линия) от сложности модели.

4. Результаты экспериментов

База данных CoverType содержит 581012 примеров и 54 независимых признака (10 непрерывнозначных и 44 булевых), задаёт задачу классификации с учителем на 7 классов при достаточно неравномерном распределении примеров по классам. Исследование возможности использования единственной нейросети для решения этой задачи показало, что для снижения доли неправильно решенных примеров до приемлемого уровня в 25% ошибок требуется нейросеть размером более 100 нейронов – данная оценка получена экстраполяцией графика на Рис.3. График на Рис.3 отражает ошибки обучения и обобщения для нейросетей разного размера, обученных на 4/5 выборки и протестированных на оставшейся доли в 1/5. Видно, что сложность модели не достигает момента начала расхождения графиков ошибок обучения и обобщения, наподобие отмеченного на Рис.2. А размер выборки достаточен для того, чтобы ошибки обучения и обобщения при любом размере сети были близки, т.е. соответствовали асимптотическому пределу Рис.1. Но нейросеть с размером в 100 нейронов и более, дающая приемлемую точность, может требовать неприемлемо медленного для практики времени срабатывания (база данных относится к задаче идентификации лесорастительного покрова мелких фрагментов территорий по спутниковым снимкам). Поэтому можно попытаться повысить точность решения путем построения коллектива экспертов, имеющих в сумме меньшее число нейронов, чем требуется для достижения нужной точности решения отдельным экспертом. А поскольку каждый эксперт будет иметь малый размер и, соответственно, значительно повышенный уровень ошибки (да, вдобавок, и более усложненные выборки, по сравнению с исходной, будут формироваться для обучения второго и последующих экспертов boosting-коллектива), то максимизация прогностических способностей каждого эксперта превращается в необходимую задачу.

Исследуем возможность построения boosting-коллектива из трех нейросетей размером в 20 нейронов каждая. Графики ошибок обучения и обобщения для первой сети-эксперта при изменении объёма обучающей выборки от 2000 до 20000 примеров с шагом в 1000 примеров даны на Рис.4. Видно, что выход на асимптоту ошибки обучения, т.е. приближение свойств обучающей выборки к свойствам генеральной совокупности и прекращение меморизации выборки, наступает при размере выборки

порядка 15000 примеров и выше. Ошибки обучения и обобщения при этом ниже таковых для 20-нейронной сети, обученной на большой выборке (Рис.3) – это значит, что некоторой нежелательной меморизации удалось избежать.

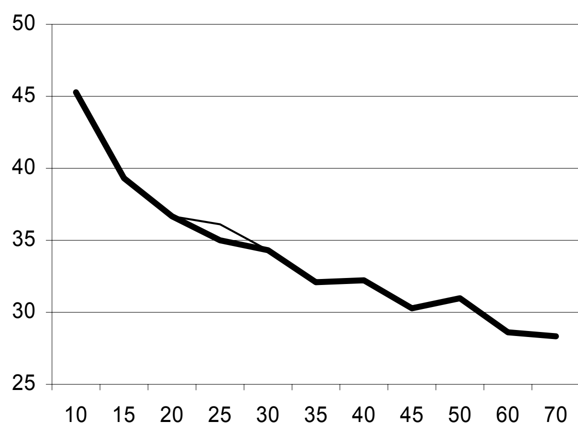


Рис.3. Ошибки обучения (толстая линия) и обобщения (тонкая линия) для нейросетей разного размера при избыточном объеме обучающей выборки (горизонтально – размер сети в нейронах, вертикально – доля неправильно решенных примеров, в %).

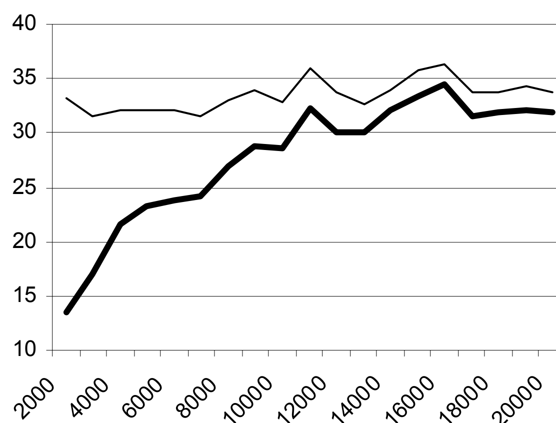


Рис.4. Ошибки обучения (толстая линия) и обобщения (тонкая линия) для первой сети-эксперта при изменении объема обучающей выборки. Объем выборки – горизонтальная ось, ошибка как % неправильно решенных примеров – вертикальная ось.

Для второй нейросети-эксперта выход на асимптоту ошибки обучения наступает тоже при размере обучающей выборки порядка 15000 примеров (Рис.5), но эта ошибка обучения гораздо большее ошибки обучения первого эксперта, что подтверждает рост сложности специально сформированной выборки, составленной из равных долей правильно и неправильно решенных первым экспертом примеров, по сравнению с генеральной совокупностью. Однако, для второй сети минимум ошибки обобщения (причем, на всей выборке, а не на построенной по тем же правилам, что и обучающая – поскольку при принятии решения две первых модели коллектива будут срабатывать всегда) достигается при меньшем объеме обучающей выборки – порядка 6000÷7000 примеров; здесь же достигаются как минимум числа тестовых примеров, для которых первый и второй эксперт дают противоречивые прогнозы, так и минимум числа неправильных ответов для тех примеров, где оба эксперта отвечают согласованно.

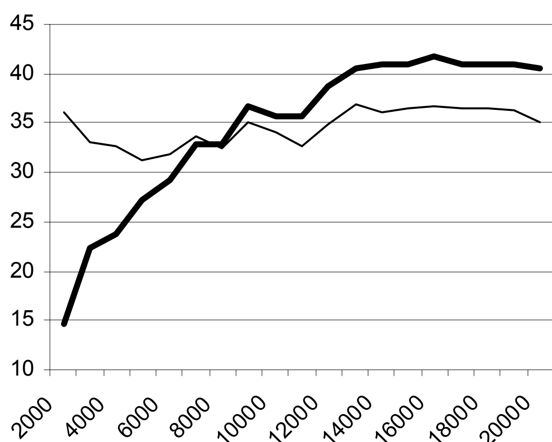


Рис.5. Ошибки обучения (толстая линия) и обобщения (тонкая линия) для второй сети-эксперта. Объем выборки – горизонтальная ось, ошибка как % неправильно решенных примеров – вертикальная ось.

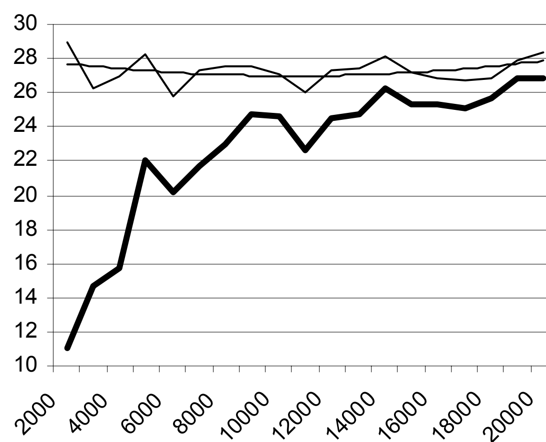


Рис.6. Ошибки обучения (толстая линия), обобщения и график квадратичного тренда над ошибкой обобщения (тонкие линии) для третьей сети-эксперта.

Третий эсперт, выходя на асимптоту ошибки обучения при обучающей выборке в 20000 примеров и выше, при аппроксимации графика ошибки обобщения квадратичным трендом указывает на минимум ошибки обобщения (на примерах, касательно которых прогноз первых двух экспертов не совпадает) при 12000 примерах в обучающей выборке (Рис.6).

Итоговый коллектив из трех экспертов, оптимизированных описанным образом, имеет ошибку распознавания на уровне 26%, что лучше уровня в 28% ошибок для единой сети из 60 нейронов, представленного на Рис.1. Уменьшение ошибки коллектива, несмотря на более высокие ошибки обобщения отдельных экспертов, обусловлено высокой правильностью ответа первых двух экспертов на тех примерах, где совпадает их прогноз – там уровень ошибки порядка 24% от числа таких примеров. Там, где первые два эксперты дают разные прогнозы, ошибка выносящего окончательный вердикт третьего эксперта тоже меньше 28% от числа таких примеров.

Предположительно, коллектив из сетей с 25-30 нейронами будет сопоставим по точности с единой нейросетью в 100 и более нейронов, для которой путем экстраполяции графика Рис.3 прогнозируется число ошибок на уровне 25%. Поскольку срабатывание третьего эксперта требуется не для каждого примера, поступающего коллективу на распознавание, то при последовательной программной реализации среднее время отклика коллектива будет на четверть или даже треть меньше, чем время срабатывания одной большой 100-нейронной сети. При сохранении же среднего времени отклика на исходном уровне общее число нейронов в нейросетях коллектива можно довести до 110-120 и еще сильнее выиграть в точности распознавания.

Ухудшение обобщающих способностей сети с ростом обучающей выборки (лучшее обобщение первого эксперта по сравнению с единой сетью, обученной на всей выборке – см. Рис.3,4; рост ошибки обобщения для третьего эксперта – см. Рис.6), противоречащее "теоретическому" графику Рис.1, возможно, иллюстрирует увеличение числа "некорректностей" (шумов, выбросов) в данных с ростом объема выборки (доля выбросов в выборке может соответствовать такой доле в генеральной совокупности, просто с ростом объёма выборки в ней растёт абсолютное число таких примеров), и значительное влияние этих некорректностей на процесс оценивания параметров модели при её обучении, т.е. переобучение. Наблюдение такого эффекта требует тщательного анализа выборок и коррекций выбросов в данных, что может ещё сильнее улучшить точность моделей коллектива.

5. Заключение

Описанные эксперименты показывают четкое проявление асимптотических свойств "кривых" обучения, при этом даже на не слишком больших выборках (несколько тысяч примеров) отсутствует необходимость многократного обучения сети при оценивании каждой точки кривой: усреднения нескольких проб не требуется, результаты имеют достаточно малый разброс для того, чтобы тренд при визуальном наблюдении идентифицировался надёжно. В итоге после построения экспериментальных графиков зависимостей качества решения от свойств моделей и размеров выборок возможно определение оптимальных объёмов обучающих выборок, сложностей моделей. Свойство boosting-алгоритма, заключающееся в специфическом последовательном формировании обучающих выборок, может приводить к сильным статистическим отличиям таких выборок от генеральной совокупности (исходной большой выборки) и друг от друга – это требует нахождения оптимальных настроек на каждом шаге boosting-алгоритма с целью максимизации как прогностических возможностей строящегося эксперта, так и всего boosting-коллектива.

Один из возможных подходов к оптимизации прогностических способностей экспертов коллектива и самого коллектива в итоге, остающийся в базовых рамках

пошаговой схемы роста коллектива и не требующий итеративной переоптимизации моделей и способа голосования, и исследован в работе.

Можно выделить 4 характеристики, требующих оптимизации при построении каждого эксперта:

1. размер обучающей выборки;
2. размер нейросети;
3. правила предобработки обучающей выборки;
4. коррекция выбросов в данных.

Две первые характеристики могут оптимизироваться путем сканирования (с некоторым шагом) набора альтернативных значений, как было сделано в данной работе для п.1. Оптимизация размера нейросети может быть проделана подобным же образом на основе данной на Рис.2 теоретической зависимости ошибок обучения и обобщения от информационной ёмкости модели. В предельном случае оптимизацию можно вести сразу в двумерном пространстве: размер выборки – размер нейросети. Оптимизация же характеристик 3,4 может быть проделана однократно для каждой сети на предельном для неё размере обучающей выборки (предполагая, что все обучающие выборки меньшего размера будут получены из такой предельной): если выборки малого размера требуют переоптимизации, то это указывает на несовпадение статистических свойств этих выборок с подобными свойствами генеральной совокупности (предельной выборки), т.е. очевидную неприемлемость таких объемов выборки.

Литература

1. *Schapire R.* The strength of weak learnability / Machine Learning, 1990. Vol.5, No.2. – pp.197-227.
2. *Freund Y., Schapire R.* Experiments with a new boosting algorithm / Proc. 13th Conf. on Machine Learning, 1993. – pp.148-156.
3. *Царегородцев В.Г.* Оптимизация предобработки данных: константа Липшица обучающей выборки и свойства обученных нейронных сетей // Нейрокомпьютеры: разработка, применение. 2003, №7. – с.3-8.
4. *Cortes C., Jackel L.D., Solla S.A., Vapnik V., Denker J.S.* Learning curves: Asymptotic values and rate of convergence / Advances in Neural Information Processing Systems 6 (1993). Morgan Kaufmann, 1994. – pp.327-334.
5. *Cortes C., Jackel L.D., Chiang W.-P.* Limits on learning machine accuracy imposed by data quality / Advances in Neural Information Processing Systems 7 (1994). MIT Press, 1995. – pp.239-246.
6. *Осовский С.* Нейронные сети для обработки информации. - М.: Финансы и статистика, 2002. – 344с.
7. *Drucker H., Schapire R., Simard P.* Improving performance in neural networks using a boosting algorithm / Advances in Neural Information Processing Systems 5 (1992), Morgan Kaufmann, 1993. – pp.42-49.
8. *Kröse B., van der Smagt P.* An introduction to neural networks. Eight edition. Univ. of Amsterdam, The Netherlands, 1996.
9. *Watanabe E., Shimizu H.* Relationships between internal representation and generalization ability in multi layered neural network for binary pattern classification problem / Proc. Int. Joint Conf. Neural Networks (IJCNN'1993), Nagoya, Japan, 1993. Vol.2. – pp.1736-1739.
10. *Gu B., Hu F., Liu H.* Modelling classification performance for large data sets: An empirical study / Lecture Notes in Computer Science, 2001. Vol.2118. – pp.317-328.